Department of Agricultural and Resource Economics
University of California, Davis

Markov-Switching Model Selection
Using Kullback-Leibler Divergence

By

Aaron Smith, Prasad A. Naik, and Chih-Ling Tsai

Working Paper No. 05-001

January, 2005

Giannini Foundation of Agricultural Economics

Markov-switching model selection using Kullback-Leibler divergence

Aaron Smith[*]
Department of Agricultural and Resource Economics
University of California, Davis

Prasad A. Naik
Graduate School of Management
University of California, Davis

Chih-Ling Tsai
Graduate School of Management
University of California, Davis
and
Guanghua School of Management
Peking University, P. R. China

Abstract

In Markov-switching regression models, we use Kullback-Leibler (KL) divergence between the true and candidate models to select the number of states and variables simultaneously. In applying Akaike information criterion (AIC), which is an estimate of KL divergence, we find that AIC retains too many states and variables in the model. Hence, we derive a new information criterion, Markov switching criterion (MSC), which yields a marked improvement in state determination and variable selection because it imposes an appropriate penalty to mitigate the over-retention of states in the Markov chain. MSC performs well in Monte Carlo studies with single and multiple states, small and large samples, and low and high noise. Furthermore, it not only applies to Markov-switching regression models, but also performs well in Markov-switching autoregression models. Finally, the usefulness of MSC is illustrated via applications to the U.S. business cycle and the effectiveness of media advertising.

[*]Corresponding author: Aaron Smith, Department of Agricultural and Resource Economics, University of California, One Shields Avenue, Davis, CA 95616.
Phone 530-752-2138; Fax 530-752-5614; E-mail adsmith@ucdavis.edu.

1. Introduction

Economic systems often experience shocks that shift them from their present state into another state; for example, nations lurch into recession, government regimes change over time, and financial markets exhibit bubbles and crashes. These states tend to be stochastic and dynamic: if they occur once, they probably recur. To capture such probabilistic state transitions over time, Markov-switching models provide an analytical framework. In economics, Markov-switching models have been used for investigating the U.S. business cycle (Hamilton 1989), foreign exchange rates (Engel and Hamilton 1990), stock market volatility (Hamilton and Susmel 1994), real interest rates (Garcia and Perron 1996), corporate dividends (Timmermann 2001), the term structure of interest rates (Ang and Bekaert 2002a), and portfolio allocation (Ang and Bekaert 2002b), among others. Outside of economics, Markov-switching models find application in diverse fields such as computational biology (e.g., Durbin et al. 1998 for gene sequencing), computer vision (Bunke and Caelli 2001), and speech recognition (Rabiner and Juang 1993).

To estimate Markov-switching models, Baum and his colleagues (Baum and Petrie 1966, Baum et al. 1970) developed the forward-backward algorithm, which was extended to encompass general latent variable models under the expectation-maximization (EM) principle (see Dempster, Laird and Rubin 1977). If the number of states in Markov-switching models is known, the EM algorithm yields consistent parameter estimates, and statistical inference proceeds via standard maximum-likelihood theory (e.g., Bickel, Ritov and Rydén 1998). If the number of states is not known, however, the likelihood ratio test to infer the true number of states breaks down because regularity conditions do not hold (see Hartigan 1977, Hansen 1992, Garcia 1998).

The number of states is often not known a priori, so we propose applying Kullback-Leibler (KL) divergence to determine it. We note that KL divergence has been used in various model selection contexts (see, e.g., Sawa 1978, Leroux 1992, Sin and White 1996, Burnham and Anderson 2002). Specifically, Akaike's information criterion (AIC, see Akaike 1973) provides an estimate of KL distance but, in Markov-switching models, it misleads the users into selecting too many states (see Section 4.2). Consequently, one fits spurious regressions in nonexistent states; this misspecification results in incorrect inclusion of variables, which reduces the accuracy of estimated parameters and lowers the precision of model forecasts. Hence, the problem of simultaneous determination of the number of states to retain in the Markov chain and the variables to include in the regression model for each retained state remains open.

The objective of this paper is to develop a new information criterion for simultaneous selection of states and variables in Markov switching models. To accomplish this goal, we obtain an explicit approximation to the KL distance for the class of Markov switching regression models. The resulting Markov switching criterion (MSC) imposes an appropriate penalty, and so it mitigates the over-retention of states in the Markov chain and alleviates the tendency to over-fit the number of variables in each state. Moreover, in Monte Carlo studies, MSC performs well in single and multiple states, small and large samples, and low and high noise. Finally, it not only applies to Markov-switching regression models, but also performs well in Markov-switching autoregression models.

We present two empirical applications of MSC to understand (a) the business cycles in the US economy and (b) the effectiveness of media advertising. In the business cycle application, based on the minimum MSC value, we retain a three-state model for US GNP growth with one recessionary state and two expansionary states. The second expansionary state

occurred mostly after 1984, and it exhibits slower growth, lower volatility, and longer duration than the first one. This finding supports the notion of "great moderation" (see Kim and Nelson 1999a, McConnell and Perez-Quiros 2000, Stock and Watson 2003). In the advertising application, MSC suggests the retention of a two-state Markov-switching model for sales and advertising of the Lydia Pinkham brand; the results reveal new insights not discernible from the standard regression model.

We organize this paper as follows. In Section 2, we describe the model structure and estimation algorithm for multiple state Markov-switching models. We derive the information criterion in Section 3 and investigate its properties and performance under various conditions in Section 4. Section 5 presents empirical applications to business cycles and media advertising. Section 6 concludes the paper by identifying avenues for future research.

2.   Estimating *N*-state Markov-switching models

We present the model structure, establish notation, and briefly describe the estimation of Markov-switching regressions, conditional on knowing the number of states *N*.

2.1   Model structure

Consider an *N*-state Markov chain. Let s_t denote an $N \times 1$ selection vector with elements $s_{ti} = 1$ or 0, according to whether the Markov chain resides in the state i $(i = 1,..., N)$. The unobserved state vector s_t evolves according to an ergodic Markov chain with the transition probability matrix

$$P = \begin{bmatrix} p_{11} & \cdots & p_{1N} \\ \vdots & p_{ij} & \vdots \\ p_{N1} & \cdots & p_{NN} \end{bmatrix}, \tag{1}$$

where $p_{ij} = pr(s_{t+1,j} = 1 \mid s_{ti} = 1)$ and $\sum_{j=1}^{N} p_{ij} = 1$ for every $i = 1,...,N$. We define the ergodic probabilities of the Markov chain by the vector $\pi = (\pi_1,...,\pi_N)'$, where $\sum_{i=1}^{N} \pi_i = 1$.

At time t, when the chain is in state i (i.e., $s_{ti} = 1$), we observe the dependent variable y_t according to the regression model

$$y_t = x_t'\beta_i + \sigma_i \varepsilon_{ti}, \tag{2}$$

where $\varepsilon_{ti} \sim N(0,1)$ is independently distributed over time $t = 1,...,T$, x_t contains K explanatory variables, and the $K \times 1$ vector β_i denotes their marginal impact when the chain is in the state i. If the chain moves to the state j, the marginal impact of exogenous variables is β_j with the corresponding level of noise σ_j^2. To capture this "switching" in regression models, we rewrite (2) as follows:

$$y_t = x_t'\beta\, s_t + \sigma\, s_t \varepsilon_t \tag{3}$$

where $\beta = (\beta_1,...,\beta_N)$, $\sigma = (\sigma_1,...,\sigma_N)$, and the selection vector s_t indicates the state at time t. The matrix β and vector σ have dimensions $K \times N$ and $1 \times N$, respectively. Equations (1) and (3), together, constitute the N-state Markov-switching regression model. When x_t includes lagged values of y_t, we obtain the N-state Markov-switching autoregression model (e.g., Hamilton 1989). Next, we describe an EM algorithm to estimate this model.

2.2   EM algorithm

Suppose we observe the complete data, including the sequences of both the observed variables $Y = \{(y_t, x_t'): t = 1,...,T\}$ and the state variables $S = \{s_t : t = 1,...,T\}$. Then the complete data log-likelihood function L_c is

$$L_c(\theta;Y,S) = L(\beta,\sigma;Y\mid S) + L(P;S)$$
$$= \sum_{t=1}^{T}\sum_{i=1}^{N} s_{ti}\log f_i(y_t;\beta_i,\sigma_i) + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ti}s_{t+1,j}\log p_{ij} + \sum_{i=1}^{N} s_{1i}\log\pi_i, \tag{4}$$

where $f_i(y_t;\beta_i,\sigma_i) = (2\pi\sigma_i^2)^{-1/2}\exp(-\tfrac{1}{2}(y_t - x_t'\beta_i)^2/\sigma_i^2)$ is the density of y_t conditional on

$s_{ti} = 1$ (see McLachlan and Peel 2000, p. 329).

In the E-step, we evaluate the expectation of L_c with respect to the unobserved latent

states S, given the observed data Y and provisional estimates of θ. Let θ^l denote the provisional

estimates at the l-th iteration, and $Q(\theta;\theta^l) = E[L_c\mid Y,\theta^l]$. Because L_c is linear in s_{ti}, $s_{ti}s_{t+1,j}$,

and s_{1i}, we obtain

$$Q(\theta,\theta^l) = \sum_{t=1}^{T}\sum_{i=1}^{N}\xi_{ti}^{(l)}\log f_i(y_t;\beta_i,\sigma_i) + \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\tau_{tij}^{(l)}\log p_{ij} + \sum_{i=1}^{N}\xi_{1i}^{(l)}\log\pi_i, \tag{5}$$

where $\tau_{tij}^{(l)} = E(s_{ti}s_{t+1,j}\mid Y,\theta^l)$ and $\xi_{ti}^{(l)} = E(s_{ti}\mid Y,\theta^l)$. To compute $(\tau_{tij}^{(l)},\xi_{ti}^{(l)})$, we apply the

forward-backward algorithm (e.g., McLachlan and Peel 2000, p. 330), which yields

$$\tau_{tij}^{(l)} = \frac{a_{ti}^{(l)} p_{ij}^{(l)} f_i(y_{t+1};\theta^l) b_{t+1,j}^{(l)}}{\sum_{i=1}^{N}\sum_{j=1}^{N} a_{ti}^{(l)} p_{ij}^{(l)} f_i(y_{t+1};\theta^l) b_{t+1,j}^{(l)}} \tag{6}$$

and

$$\xi_{ti}^{(l)} = \sum_{j=1}^{N}\tau_{tij}^{(l)}. \tag{7}$$

The "forward" probabilities a_{ti} are given by the forward recursion

$$a_{t+1,i}^{(l)} = \left(\sum_{j=1}^{N} a_{tj}^{(l)} p_{ij}^{(l)}\right) f_i(y_{t+1};\theta^l), \tag{8}$$

and the "backward" probabilities b_{tj} are given by the backward recursion

$$b_{tj}^{(l)} = \sum_{i=1}^{N} p_{ij}^{(l)} f_i(y_{t+1};\theta^l) b_{t+1,i}^{(l)}. \tag{9}$$

We initialize these recursions by setting $a_{1i} = \pi_i^{(l)} f_i(y_1;\theta^l)$ and $b_{Tj} = 1$, where

$\pi^{(l)} = (\pi_1^{(l)},...,\pi_N^{(l)})'$ is the principal eigenvector of $P^{(l)}\pi = \pi$.

In the M-step, we maximize $Q(\theta,\theta^l)$ with respect to $\theta = vec(\beta,\sigma,P)$ to obtain the closed form estimates for the $(l+1)$-th iteration:

$$\beta_i^{(l+1)} = (X'W_i^{(l)}X)^{-1}X'W_i^{(l)}y, \tag{10}$$

$$(\sigma_i^{(l+1)})^2 = (y - X\beta_i^{(l+1)})'W_i^{(l)}(y - X\beta_i^{(l+1)})/T_i^{(l)}, \tag{11}$$

and

$$p_{ij}^{(l+1)} = \frac{\sum_{t=1}^{T-1}\tau_{tij}^{(l)}}{\sum_{t=1}^{T-1}\xi_{ti}^{(l)}}, \tag{12}$$

where $X = (x_1,...,x_T)'$, $y = (y_1,...,y_T)'$, $W_i^{(l)} = diag(\xi_i^{(l)})$, $\xi_i^{(l)} = (\xi_{1i}^{(l)},...,\xi_{ti}^{(l)},...,\xi_{Ti}^{(l)})'$, and $T_i^{(l)} = tr(W_i^{(l)})$. Using the provisional estimates θ^l, we obtain the new estimates $\theta^{(l+1)} = vec(\beta^{(l+1)},\sigma^{(l+1)},P^{(l+1)})$ via the equations (10) through (12). We iterate the E- and M-steps until the absolute difference $|\theta^{(l+1)} - \theta^{(l)}|$ decreases below a preset tolerance. The resulting vector $\hat{\theta} = vec(\hat{\beta},\hat{\sigma},\hat{P})$ converges to the maximum likelihood estimates, which are consistent and asymptotically normal (Bickel, Ritov and Rydén, 1998). For finite sample properties, see Psaradakis and Sola (1998). We close this section with two remarks.

Remark 1. We enhance the stability of this algorithm as follows. First, to avoid singularities in the likelihood function and reduce the chance of spurious local maxima, we follow Hathaway's (1985) suggestion to set a lower bound on the relative variances across states. Second, to prevent underflow of forward probabilities in (8), for each t and $i = 1,..., N$, we follow Leroux and Puterman's (1992) recommendation to scale a_{ti} with a constant r such that $10^{-r}\sum_{i=1}^{N}a_{ti}$ lies between 0.1 and 1.0 and then multiply it by 10^{-r}. Because a_{ti}, appears in both the numerator and denominator of (6), the value of $\tau_{tij}^{(l)}$ does not change. Similarly, we prevent

underflow of backward probabilities in (9).

Remark 2. This EM algorithm enables the estimation of Markov-switching models with many observations because the forward-backward method is linear in T. Furthermore, because both the E- and M-steps are available in closed form, the EM algorithm is robust to numerical uncertainties encountered by quasi-Newton methods. For example, Hamilton (1990, pp. 40-41) notes that "…methods that seek to approximate the sample Hessian can easily go astray …By contrast, the EM algorithm by construction finds an analytical interior solution to a particular subproblem." Nonetheless, like quasi-Newton methods, the EM does not guarantee convergence to global maxima (see McLachlan and Krishnan 1997, p. 34). Finally, the EM algorithm can also be used to obtain Bayesian modal values by augmenting the expected complete data likelihood with the logarithm of prior density; see Dempster, Laird and Rubin (1977, p. 6) for this connection between EM and Bayesian analysis and Kim and Nelson (1999b, Ch. 9) for implementation in Markov-switching models.

3. Deriving Markov-switching criterion

In the above estimation, the number of states N is assumed known, which need not be the case in practice. To determine the number of states, we approximate the true data generating process (DGP) using several candidate models, quantify the information loss between the DGP and each candidate model, and then choose the model that entails the minimum expected information loss (e.g., Burnham and Anderson 2002). Specifically, let $g(Y^*)$ denote the probability density function of the DGP and $f(Y^*;\theta)$ be the density function for a candidate model, where Y^* represents the data used for evaluating the model. As in Sawa (1978) and Sin and White (1996), we quantify information loss using the Kullback-Leibler (KL) divergence,

which is defined as

$$d_{KL}(g,f;\theta) = E_{Y^*}\left[\log\frac{g(Y^*)}{f(Y^*;\theta)}\right],$$ (13)

where $d_{KL} \geq 0$, and $E_{Y^*}(\cdot)$ denotes the expectation with respect to the data generating density g.

Equation (13) measures the divergence between the two densities g and f, indicating the information loss entailed when we approximate the DGP using a candidate model. Recently, Zellner (2002, p. 43) interprets d_{KL} as the difference in expected log heights of the two densities; for other divergence measures, see Rényi (1970) or Linhart and Zucchini (1986, p. 18).

The information loss in (13) depends on the model parameters θ. In practice, we evaluate (13) at $\hat{\theta}$ obtained by fitting the candidate model f with the observed sample Y. To remove the dependence of (13) on the particular sample Y, we adopt Akaike's (1985) approach to average d_{KL} across different independent samples Y drawn from the same DGP and choose a model that minimizes the expected information loss:

$$\bar{d}_{KL}(g,f;\hat{\theta}) = E_Y\left(E_{Y^*}\left[\log\frac{g(Y^*)}{f(Y^*;\hat{\theta})}\right]\right)$$

$$= E_Y(E_{Y^*}[\log(g(Y^*))]) - E_Y(E_{Y^*}[\log(f(Y^*;\hat{\theta}))]),$$

where $E_Y(\cdot)$ indicates expectation with respect to the density g which generates the estimation sample, Y.

Because $E_Y(E_{Y^*}[\log(g(Y^*)]])$ remains invariant across all candidate models (i.e., constant across different choices of f), it is sufficient to select the model that minimizes

$$\tilde{d}_{KL} = \tilde{d}_{KL}(g,f;\hat{\theta}) = -2E_Y(E_{Y^*}[\log(f(Y^*;\hat{\theta}))]),$$ (14)

where the dependence on g arises from the double expectation, and the multiplication by two is

for convenience. To derive an estimator for $\tilde{d}_{KL}$, we consider the Markov-switching regression model in (1) and (3) in which x_t does not contain lagged dependent variables.   In the Appendix, we simplify (14) and obtain the *Markov-switching criterion*,

$$MSC = -2\log(f(Y,\hat{\theta})) + \sum_{i=1}^{N} \frac{\hat{T}_i(\hat{T}_i + \lambda_i K)}{\delta_i \hat{T}_i - \lambda_i K - 2}, \tag{15}$$

where $\log(f(Y,\hat{\theta}))$ is the maximized log-likelihood value, $\hat{T}_i = tr(\hat{W}_i)$, $\hat{W}_i = diag(\hat{\xi}_{1i},...,\hat{\xi}_{Ti})$, $\delta_i = E[\pi_i^*/\hat{\pi}_i]$, $\lambda_i = E[(\pi_i^*/\hat{\pi}_i)^2]$, and π_i^* is the *i*-th element of the principal eigenvector of $P^*\pi^* = \pi^*$ for the best estimates $\theta^* = vec(\beta^*,\sigma^*,P^*) = \arg\min_\theta E_{Y^*}[-\log f(Y^*;\theta)]$.   The subsequent remarks elaborate the properties of MSC and its implementation in practice.

Remark 3. The first term of MSC measures the lack of fit; its second term imposes a penalty for including redundant states and variables. Thus, MSC balances the trade-off between improving a model's fit to the data and achieving parsimony of the fitted model. To select the candidate model, we compute (15) for varying choices of states and variables (*N*, *K*) and retain the one that attains the smallest value.

Remark 4. In regression models without Markov switching, MSC is equivalent to both Hurvich and Tsai's (1989) criterion in finite samples and Akaike's (1973) criterion in large samples. Specifically, in regression models, $N = \delta = \lambda = 1$, and so $MSC_{N=1} = -2\log(f(Y,\hat{\theta})) + T(T+K)/(T-K-2)$, which equals Hurvich and Tsai's (1989, p. 300) AIC_C criterion. Furthermore, by subtracting T from $MSC_{N=1}$, we obtain $-2\log(f(Y,\hat{\theta})) + 2(K+1)\{T/(T-K-2)\}$, which approaches Akaike's (1973) $AIC = -2\log(f(Y,\hat{\theta})) + 2(K+1)$ in large samples. Thus, the proposed MSC generalizes the applicability of these criteria to *N*-state Markov-switching regression models.

Remark 5.   When $N > 1$, MSC imposes penalty through $\delta_i = E[\pi_i^* / \hat{\pi}_i]$ and $\lambda_i = E[(\pi_i^* / \hat{\pi}_i)^2]$. Because the distribution of $\hat{\pi}_i$ is not known, to implement MSC, we investigate the behavior of $\bar{\delta}_i = E[\pi_i^* / \bar{\pi}_i]$ and $\bar{\lambda}_i = E[(\pi_i^* / \bar{\pi}_i)^2]$, where $\bar{\pi}_i = T^{-1} \sum_{t=1}^{T} s_{ti}$ and $E[\bar{\pi}_i] = \pi_i^*$. For $\bar{\delta}_i$, we invoke Jensen's inequality to obtain $\bar{\delta}_i = \pi_i^* E[1 / \bar{\pi}_i] \geq \pi_i^* / E[\bar{\pi}_i] = 1$. In other words, a lower bound for $\bar{\delta}_i$ is unity, which yields a larger value of MSC than would result from any other $\delta_i > 1$. For $\bar{\lambda}_i$, we applied Gabriel's (1959) formula for the distribution of $\bar{\pi}_i$ to compute $\bar{\lambda}_i$ for various $N{\times}N$ transition matrices P. These computations indicated that $\bar{\lambda}_i$ is an increasing function of the number of states N. Using these results, we set $\delta_i = 1$ and $\lambda_i = 1$, N, and N^2 to implement MSC. In Section 4, Monte Carlo simulations show that MSC with $\delta_i = 1$ and $\lambda_i = N$ performs satisfactorily.

Remark 6. The application of MSC in (15) is not specific to the EM algorithm; it can be used in conjunction with other estimation approaches. For example, one could obtain $\log(f(Y, \hat{\theta}))$ via quasi-Newton methods and find $\hat{T}_i$ using the smoother in Hamilton (1990) or Kim (1994). Thus, the value of MSC in (15) can be computed to determine states and variables jointly.

Remark 7. We obtained average $\bar{d}_{KL}$ to remove dependence of (13) on the estimation sample *Y*. Alternatively, we can consider the possibility of averaging by using a posterior density for θ and a predictive density for Y^*. This approach may provide better results in small samples, an issue that needs further investigation.

Remark 8. Bates and Granger (1969) and Leamer (1978) suggest combining multiple models rather than selecting the single best one. To this end, Burnham and Anderson (2004,

pp. 269-274) recommend computing $\Delta_k = MSC_k - MSC_{\min}$ for each model f_k relative to the model that yields the minimum MSC value, and then use the weights $w_k = \dfrac{\exp(-0.5\Delta_k)}{\sum\limits_k \exp(-0.5\Delta_k)}$ to conduct multi-model inference. Furthermore, to assess degrees of confidence in alternative models, Burnham and Anderson (2002, p. 170) offer the following guidelines: Δ_k between 0-2 indicates a substantial empirical support for the model f_k; Δ_k between 4-7 suggests considerably less support; $\Delta_k > 10$ implies essentially no empirical evidence in favor of that model (also see Raftery (1996, p. 252) for guidelines when using Bayes factors). Finally, alternative approaches for incorporating model uncertainty include forecast combinations (Timmermann 2005), Bayesian model averaging (e.g., Hoeting et al. 1999), frequentist model averaging (Hjort and Claeskens 2003), and adaptive mixing of methods (Yang 2001).

Remark 9. We note that model comparisons based on AIC are asymptotically equivalent to those based on Bayes factors when prior information is as precise as the likelihood (Kass and Raftery 1995, p. 790). When prior information is small relative to the information contained in data, the Bayesian information criterion (BIC) tends to select models with highest posterior probability. In investigating the number of states to retain in Markov-switching autoregressive models, Psaradakis and Spagnolo (2003, p. 246) conclude that BIC tends to underestimate the number of states. We encourage further research to investigate such comparisons using the proposed MSC.

Remark 10. Here we elucidate the theoretical justification for using Kullback-Liebler divergence in model selection. In information theory, Shannon's (1948) entropy is defined as $-\sum\limits_x p(x)\log(p(x))$ for a discrete random variable with probability mass function p(x). Generalizing Shannon's entropy to two continuous density functions *g* and *f*, Kullback and

Leibler (1951) quantify "information" by defining $d_{KL} = \int g(x)\log(g(x)/f(x))dx$ and by

connecting it to R. A. Fisher's notion of sufficient statistics. Akaike (1973, 1985) not only

extends Kullback-Leibler information to quantify expected information loss

(i.e., $-E[E[\log(f(x)]]]$), but also deepens the connection with likelihood theory (see deLeeuw

1992) by showing that (a) the maximized log-likelihood value is a biased estimate of expected

information loss, and (b) the magnitude of asymptotic bias equals the number of estimable

parameters in the approximating model f. These theoretical findings furnish the justification for

using KL divergence as a bridge between estimation theory and model selection, thereby

unifying them under a common optimization framework (for further details, see Burnham and

Anderson 2004, p. 268).

4.   Monte Carlo studies

Here we describe the simulation settings as well as the model selection procedure, and

then we present Monte Carlo results to illustrate the properties and performance of MSC. We

also explore the applicability of MSC to Markov-switching autoregression models.

4.1   Simulation settings and model selection procedure

We investigate the following five settings:

(i) *Markov-switching regression*: The true model consists of two states ($N^0 = 2$) and three

variables in each state including an intercept. The true regression coefficients are

$\beta^0 = (\beta_1^0, \beta_2^0)$, where $\beta_1^0 = (1, 2, 3)'$ and $\beta_2^0 = (4, 3, 2)'$. The explanatory variables are

stored in the $T \times 3$ matrix X^0, whose first column equals one and second and third

columns are randomly drawn from a standard normal distribution. The $N^0 \times 1$ state

variable s_t^0 is a Markov chain with transition probabilities: $p_{ii}^0 = 0.95$ and $p_{ij}^0 = 0.05$ for each i, j = 1, 2. We obtain the dependent variable using the model in (3), $y_t = (x_t^0)' \beta^0 s_t^0 + \sigma^0 s_t^0 \varepsilon_t^0$, where x_t^0 denotes the t-th row of X^0, t = 1, 2, …, T = 250, $\varepsilon_t^0 \sim N(0,1)$, and $\sigma^0 = (\sigma_1^0, \sigma_2^0) = (0.5, 0.5)$.

In each state, we consider five candidate variables, which are stored in the matrix X of dimension $T \times 5$. The first three columns of X are the same as X^0, and we randomly draw the last two columns from the standard normal distribution. We consider four candidate states (i.e., N = 1, …, 4), and the candidate regression models include up to five variables from X in a sequentially nested fashion. Thus, we have 20 possibilities (4 states by 5 variables) from which to choose the true model.

(ii) *Markov-switching regression with small sample and high noise:* We consider two variations from the settings in (i). First, to study small sample performance, we conduct the above simulations using T = 100. Second, we set $\sigma_i^0 = 1$ for both T = 100 and 250 to understand the effect of a higher noise level.

(iii) *Markov-switching autoregression:* We conduct the simulation in (i) for autoregressive models, where the t-th rows of X^0 and X contain (1, y_{t-1}, y_{t-2}), and (1, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}), respectively, for t = 5, 6, …, T. The true coefficients, $\beta_1^0 = (1, 0.2, 0.3)'$ and $\beta_2^0 = (3, 0.5, 0.2)'$, satisfy the stationarity condition.

(iv) *Markov-switching autoregression with small sample and high noise:* Analogous to (ii), we investigate two variations from the settings in (iii).

(v) *Single state model:* We investigate the case with N^0 = 1 to examine whether MSC leads to spurious Markov-switching structure when the true model is a standard regression. For

fixed regressors, we use $\beta^0 = (1, 2, 3)'$; for autoregression, $\beta^0 = (1, 0.2, 0.3)'$.

We conduct 1000 repetitions in each of the above settings to assess how often MSC selects the true model.

We employ the following model selection procedure for each of the 20 state-variable combinations $\{(N, K): N = 1, \ldots, 4, K = 1, \ldots, 5\}$. First, we choose initial parameter values using the K-means method (MacQueen 1967) to classify observations in the matrix (y, X) into N states. Then, we apply the EM algorithm to estimate the Markov-switching regression model. Next, we compute MSC in (15). We also constrain the term $(\delta_i \hat{T}_i - \lambda_i K - 2)$ in (15) to exceed unity in each realization to ensure positive penalty. Finally, we select the model that yields the smallest MSC value across all the 20 state-variable combinations.

4.2   Monte Carlo results

Here we present one figure and five tables to illustrate the accuracy and performance of MSC. In addition, we substantiate the claim that AIC overestimates the number of states in Markov-switching models.

Accuracy of MSC. We assess the accuracy of MSC by computing its proximity to the true KL distance. To this end, we estimate the true KL distance $\tilde{d}_{KL}$ in (14) using the three steps: (a) randomly draw an estimation sample Y to obtain the EM estimates $\hat{\theta}$; (b) draw a holdout sample Y^* to evaluate $\log(f(Y^*; \hat{\theta}))$; (c) perform 100 repetitions with different holdout samples Y^*'s to estimate $E_{Y^*}[\log(f(Y^*; \hat{\theta}))]$. We repeat the steps (a)-(c) 100 times for different estimation samples Y to evaluate the double-expectation in (14).

Figure 1 presents the proximity plots for the MSC values from (15) using $\lambda_i = N$ for the setting (i) in Section 4.1. Panel A presents the results for state selection. It shows that both MSC

and $\widetilde{d}_{KL}$ achieve their minimum at the true number of states, i.e., $N^0 = 2$. Furthermore, MSC and $\widetilde{d}_{KL}$ are close when $N \leq N^0$, while MSC exceeds $\widetilde{d}_{KL}$ when $N > N^0$. In other words, MSC approximates $\widetilde{d}_{KL}$ reasonably well and imposes a larger penalty when the number of states exceeds those in the data generating process. This larger penalty mitigates overestimation of the number of states. Panel B, which presents the results for variable selection, depicts that MSC and $\widetilde{d}_{KL}$ are uniformly close. Thus, for the purposes of model selection, the proposed MSC reasonably approximates the Kullback-Leibler distance.

Performance of MSC. We investigate the simultaneous selection of states and variables in Markov-switching regression models (see the setting (i) in Section 4.1). We assess the performance of a criterion by the relative frequency of selecting various states and variables, while the measure of accuracy is how often the criterion selects the correct number of states that were used in the DGP. Table 1 reports the frequency of correct state and variable selection using MSC with $\lambda = 1$, N and N^2. (Note that the subscript i on λ is suppressed in the rest of the paper.) For $\lambda = 1$, Panel A shows that incorrect model selection is asymmetric. Specifically, the zeros in Panel A reveal that $\text{MSC}_{\lambda=1}$ never underestimates the number of states or variables. But, $\text{MSC}_{\lambda=1}$ correctly selects two states 360 times and three variables 666 times out of 1000 occasions. Consequently, the joint frequency of selecting the correct states and variables is only 30.9%. Despite this unsatisfactory performance, we note that the *conditional* frequency of variable selection $(0, 0, \frac{309}{360}, \frac{29}{360}, \frac{22}{360})$ is satisfactory. This finding can be explained using Panel B of Figure 1, which shows that MSC estimates the true KL distance accurately when the number of states is known. More importantly, this finding underscores the insight that the model selection performance can be improved if we determine the true states accurately. To this end,

we investigate the performance of MSC with $\lambda = N$ and N^2 as stated in Remark 5.

For $\lambda = N$, Panel B indicates a marked improvement in model selection performance. Specifically, $MSC_{\lambda=N}$ correctly selects the two-state model in each of the 1000 realizations. We explain this improvement using Panel A of Figure 1, which exhibits that $MSC_{\lambda=N}$ imposes larger penalty than the KL distance, thus mitigating the tendency to fit too many states. Moreover, we find diminishing returns to further increases in penalty via $\lambda = N^2$ because performance improves marginally beyond that due to $MSC_{\lambda=N}$ (see Panel C in Table 1).

Table 2 demonstrates the robustness of these findings via the simulation setting (ii). When we increase the noise level from $\sigma_i^0 = 0.5$ to 1, the performance of $MSC_{\lambda=1}$ further deteriorates. The joint frequency of correctly selecting both the states and variables decreases from 309 to 124. In contrast, $MSC_{\lambda=N}$ and $MSC_{\lambda=N^2}$ perform well, as evidenced by the small decrease in the joint frequency from 992 to 981 and from 1000 to 998, respectively. In other words, these small decreases indicate that the performance of both the criteria do not deteriorate substantially as the noise level increases. We observe qualitatively similar findings when the sample size decreases from $T = 250$ to 100. It is worth noting that $MSC_{\lambda=N}$ is less sensitive to noise level in small samples than $MSC_{\lambda=N^2}$. Specifically, as the noise level increases for $T = 100$, the joint selection frequency of $MSC_{\lambda=N}$ decreases by 4.6% (from 951 to 907) compared to 13.9% for $MSC_{\lambda=N^2}$ (from 861 to 741). In other words, $MSC_{\lambda=N}$ outperforms $MSC_{\lambda=1}$ and $MSC_{\lambda=N^2}$ when both the sample size is small and the signal is weak.

We repeat the above analyses for the Markov-switching *autoregression* models described in the setting (iii). Table 3 reports the joint selection frequency by MSC with $\lambda = 1$, N, and N^2 in 1000 realizations. As before, incorrect model selection is asymmetric; $MSC_{\lambda=1}$ never understates

the number of states and seldom underestimates the number of variables. $MSC_{\lambda=N}$ outperforms $MSC_{\lambda=1}$ with 979 correct selections out of 1000 occasions (see Panel B in Table 3). This superior performance is due to the penalty imposed by $MSC_{\lambda=N}$, which mitigates the tendency to fit excessive states. We can marginally improve this performance from 979 to 984 by using a stronger penalty via $\lambda = N^2$ (compare Panels B and C in Table 3).

Table 4 shows that these findings are robust to various scenarios in the setting (iv). As the noise level increases in large samples, $MSC_{\lambda=1}$ performs poorly, whereas $MSC_{\lambda=N}$ and $MSC_{\lambda=N^2}$ perform satisfactorily as evidenced by smaller decreases in the joint frequency. We obtain qualitatively similar results for the small sample case. Moreover, $MSC_{\lambda=N}$ is less sensitive to the noise level in small samples than $MSC_{\lambda=N^2}$; for example, the correct selection frequency of $MSC_{\lambda=N}$ decreases by 46% (from 744 to 402) compared to 99.4% for $MSC_{\lambda=N^2}$ (from 171 to 1). Thus, $MSC_{\lambda=N}$ outperforms $MSC_{\lambda=1}$ and $MSC_{\lambda=N^2}$ when both the sample size is small and the signal is weak.

Single-state model. While MSC detects Markov switching when it does exist, can MSC reject Markov switching when it does not exist? To answer this question, we examine the setting (v) and use MSC to select the number of states (but not variables). In Table 5, Panels A and B show the correct selection frequency for the fixed regressor and autoregression settings, respectively. We find that $MSC_{\lambda=1}$ performs poorly regardless of the noise level or the sample size. However, the last two columns indicate that $MSC_{\lambda=N}$ and $MSC_{\lambda=N^2}$ correctly select a single-state model more than 90% of the occasions. Thus, $MSC_{\lambda=N}$ and $MSC_{\lambda=N^2}$ do not yield spurious Markov-switching structure when the true model is a standard regression.

We close this section by substantiating the claim in the *Introduction* that the AIC-based

estimate of KL divergence retains too many states and variables.   We compute AIC $= -2\log f(y,\hat{\theta}) + 2d$, where $d = (NK + N^2)$ denotes the number of free parameters in θ. For the sake of illustration, we use the low noise and large sample setting (ii), which is favorable for AIC. Table 6 reveals that AIC selects more states and variables than in the DGP and that the correct joint selection frequency is only 48.1%. Thus, by using AIC in practical applications, users stand about equal chance to retain a correct or an incorrect model; when it is the latter, they would fit spurious regressions in non-existent states. We next present two empirical examples to illustrate the usefulness of $\text{MSC}_{\lambda=N}$ in practice.

5.   Empirical examples

We first study the business cycle in the US economy and then the effectiveness of media advertising.

5.1   U.S. real GNP growth

Hamilton (1989) was first to formulate the Markov-switching autoregression model to capture business cycles in real GNP. In his formulation, the mean GNP growth rate switches between two states: recessions and expansions. Hansen (1992) extends this model to allow both the mean growth rate and the autoregressive coefficients to switch between states. We study this extended model, which is given by equations (1) and (3), where $x_t = (1, y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4})'$ and y_t is quarterly real GNP growth in chained 1996 dollars. We use seasonally adjusted data that span the period 1947:1 through 2002:4 (see http://www.bea.doc.gov). We exclude 16 quarterly observations (1999:1 to 2002:4) from the estimation sample and use these excluded observations to evaluate one-quarter-ahead forecasts. The estimation sample comprises $T = 203$ observations because we also exclude 5 observations for computing the growth rate and the initial lagged

values.

We apply the EM algorithm described in Section 2 to these data, and consider various state-variable combinations (N, K), where $N = 1, \ldots, 4$ and $K = 1, \ldots, 5$. We estimate 20 different N-state Markov-switching autoregression models and compute the two estimates of KL divergence: AIC and $\text{MSC}_{\lambda=N}$. Based on the minimum AIC value, we would select a model with $N^* = 4$ and $K^* = 5$, which is the largest model in this set of 20 candidate models. This finding is consistent with the simulation evidence (see Table 6), which reveals AIC's tendency to select more states and variables than necessary.

On the other hand, the minimum value of $\text{MSC}_{\lambda=N}$ yields $N^* = 3$ and $K^* = 1$, indicating the retention of the three-state model with no autoregressive lags (i.e., intercepts only). Table 7 reports the parameter estimates for this retained model, which identifies one recessionary state ($i = 1$) and two expansionary states ($i = 2, 3$). The estimated decline in real GNP during recessions is -0.10% per quarter; the mean growth rates during the two expansion states are 1.50% and 0.85% per quarter.

In Figure 2, we present the estimated smoothed probability sequence $\hat{\xi}_i = (\hat{\xi}_{1i}, \ldots, \hat{\xi}_{Ti})'$ based on (7) and overlay it with the recessionary periods (in gray bars) noted by the National Bureau of Economic Research. Panel A shows that the estimated probability of recession reasonably matches the actual recessions. Panels B and C display the two types of expansions. The first type occurred exclusively before 1984, while the second occurred mostly during the 80s and the 90s. Because $\hat{\sigma}_3 = 0.42 < \hat{\sigma}_2 = 0.91$, the recent expansionary state ($i = 3$) exhibits lower volatility than the previous one ($i = 2$). This finding supports the phenomenon of great moderation — first discovered by Kim and Nelson (1999a) and McConnell and Perez-Quiros (2000) — which is characterized by a reduction in the variance of economic growth since 1984.

We compare the forecasting performance of this retained model to that of a benchmark model that specifies Ln(GNP) as a random walk with constant drift. Over the period 1999-2002, the mean squared forecast errors are 0.351 and 0.433 for the retained model and the random walk model, respectively. In addition, the mean absolute forecast error was 0.539 for the retained model and 0.546 for the random walk. The retained three-state Markov-switching model performs well because it adapts to the recession in 2001, whereas the random walk model does not (see Figure 2).

5.2   Advertising effectiveness

In marketing, brand managers commonly use the advertising model, $y_t = \beta^{(0)} + \beta^{(1)} z_t + \beta^{(2)} y_{t-1} + \varepsilon_t$, to determine the effectiveness of advertising (Bucklin and Gupta 1999, p. 262), where y_t denotes brand sales at time t, z_t represents advertising spending, and ε_t is the normal error term. The coefficient $\beta^{(1)}$ measures the effectiveness of current advertising; the coefficient $\beta^{(2)}$, known as the carryover effect, captures the cumulative impact of past advertising reflected in the attained sales y_{t-1} (see, e.g., Palda 1964, p. 13). We extend this advertising model by incorporating regime shifts so that the parameter vector $\beta_i = (\beta_i^{(0)}, \beta_i^{(1)}, \beta_i^{(2)})'$ is specific to each regime $i = 1, \ldots, N$. This extension marks the first application of Markov-switching models in the advertising literature (see Feichtinger, Hartl and Sethi 1994, Mantrala 2002, Naik and Raman 2003).

We apply this extended model to Lydia Pinkham company's annual sales and advertising data from 1914 through 1960 (Palda 1964). This classic data set exhibits a few unique features: relatively stable product design during this period; advertising primarily affects sales, given the absence of channel members or sales force; and the lack of close competitors. These market conditions comport with the above advertising model. Furthermore, after the second World War

ended, Lydia Pinkham management demonstrated the product's efficacy to the Federal Trade Commission (FTC), which permitted them to make stronger claims in their advertising copy. Moreover, they switched from pure newspaper advertising to a mix of multiple media, which comprised newspaper, magazine, radio, and even television. (See Palda 1964, pp. 25-26 for details.)

Given these changes in market conditions, we consider the possibility of a distinct post-war regime(s) by estimating various Markov-switching models with state-variable combinations (N, K), for $N = 1, \ldots, 4$ and $K = 1, \ldots, 3$. Then we compute AIC and $\text{MSC}_{\lambda=N}$ for each combination. AIC selects a model with $N^* = 3$ states, which, given the simulation results in Table 6, is likely to be more than necessary. In contrast, $\text{MSC}_{\lambda=N}$ retains two states (i.e., $N^* = 2$). The smoothed probabilities $\hat{\xi}_1 = (\hat{\xi}_{11}, \ldots, \hat{\xi}_{T1})'$ indicate that the first state persisted from 1914 through 1945, whereas the second state lasted from 1946-1960. This regime switch coincided with the FTC's approval of stronger copy and the beginning of multiple media spending.

Table 8 shows the different estimates of advertising effectiveness and carryover effects for the pre- and post-war regimes. Specifically, advertising is more effective in the post-war era ($\hat{\beta}_2^{(1)} = 1.17 > \hat{\beta}_1^{(1)} = 0.43$) due to stronger copy and multiple media. In addition, the carryover effect is smaller in the post-war era ($\hat{\beta}_2^{(2)} = 0.27 < \hat{\beta}_1^{(2)} = 0.53$), given the shorter duration for the impact of past advertising to accumulate. Thus, these new findings are not discernible from the standard regression model of advertising.

6.   Concluding remarks

Markov-switching regression models provide an analytical framework to study both shifts in regimes and the differential impact of explanatory variables across regimes (or states).

In this paper, we investigate the problem of selecting an appropriate Markov-switching model by applying the principle of minimum Kullback-Leibler divergence. Specifically, we derive a new Markov-switching criterion, MSC, to jointly determine the number of states and variables to retain in the model. We find that MSC performs well not only in regression and autoregression models, but also in single and multiple states, small and large samples, and low and high noise. Furthermore, it provides valuable insights in empirical applications. For example, it identifies three states — one recession and two expansions — in real GNP data; the second expansion exhibits slower growth, lower volatility and longer duration than the first one, an insight that is consistent with the notion of "great moderation" (Kim and Nelson 1999a, McConnell and Perez-Quiros 2000, Stock and Watson 2003). In the advertising study, MSC enables brand managers to detect shifts in market conditions and to estimate advertising and carryover effects specific to every identified market condition.

We conclude by identifying four avenues for further research. The first one is to extend MSC to the "mixed" switching regression case, where some coefficients do not change across states, while the others do. The second is to allow different explanatory variables in each regime. The third avenue is to incorporate non-linearity in (2) via the single-index model (e.g., Horowitz 1998); see Naik and Tsai (2001) for model selection in the single-state case. Finally, we encourage research to investigate model selection for periodic regime-switching models (Ghysels, McCulloch and Tsay 1998) and state space models with time-varying coefficients (Kim and Nelson 1999b, Naik, Mantrala and Sawyer 1998). We believe that such efforts would enhance the usefulness of Markov-switching regression models.

Acknowledgements

Appendix: Derivation of Markov-switching criterion

Let $\hat{\theta} = vec(\hat{\beta}, \hat{\sigma}, \hat{P})$ be the MLE of θ computed from a realization Y that is independent

of Y^*.   In addition, let $\theta^* = vec(\beta^*, \sigma^*, P^*) = \arg\min_\theta E_{Y^*}[-\log f(Y^*; \theta)]$ and S^* denote a

realization from a Markov chain of dimension N with transition probability matrix P^*. Then the

average KL information loss is

$$\tilde{d}_{KL} = -2E_Y E_{Y^*}\left(\log(f(Y^*; \hat{\theta}))\right)$$

$$= -2E_Y E_{Y^*,S^*}\left(\log f(Y^*; \hat{\theta})\right)$$

$$= -2E_Y E_{Y^*,S^*}\left(\log f(Y^* \mid S^*; \hat{\beta}, \hat{\sigma}) + \log f(S^*; \hat{P}) - \log f(S^* \mid Y^*; \hat{\theta})\right)$$

$$= -2E_Y E_{Y^*,S^*}\left(\sum_{t=1}^{T}\sum_{i=1}^{N} s_{ti}^* \log f_i(y_t^*; \hat{\beta}_i, \hat{\sigma}_i)\right) - 2E_Y E_{Y^*,S^*}\left(\sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ti}^* s_{t+1,j}^* \log \hat{p}_{ij} + \sum_{i=1}^{N} s_{1i}^* \log \hat{\pi}_i\right)$$

$$+ 2E_Y E_{Y^*,S^*}\left(\sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ti}^* s_{t+1,j}^* \left(\log \hat{\tau}_{tij} - \log \hat{\xi}_{ti}\right) + \sum_{i=1}^{N} s_{1i}^* \log \hat{\xi}_{1i}\right)$$

$$= -2E_Y E_{Y^*,S^*}\left(\sum_{t=1}^{T}\sum_{i=1}^{N} s_{ti}^* \log f_i(y_t^*; \hat{\beta}_i, \hat{\sigma}_i)\right) - 2E_Y E_{Y^*}\left(R(N, Y^*, \hat{\theta})\right), \qquad (A.1)$$

where $E_{Y^*,S^*}(\cdot)$ indicates the expectation under the joint density of (Y^*, S^*), and

$$R(N, Y^*, \hat{\theta}) \equiv \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} \tau_{tij}^* \log \hat{p}_{ij} + \sum_{i=1}^{N} \xi_{1i}^* \log \hat{\pi}_i - \sum_{t=1}^{T-1}\sum_{i=1}^{N}\sum_{j=1}^{N} \tau_{tij}^* (\log \hat{\tau}_{tij} - \log \hat{\xi}_{ti}) - \sum_{i=1}^{N} \xi_{1i}^* \log \hat{\xi}_{1i} .$$

We assume $\hat{\pi}_i > 0$ almost surely, i.e., the estimated probability that the process visits each state

is positive. Also, note that all expectations are conditional on x_t, $t = 1, 2, \ldots, T$. The first term in

(A.1) is

$$-2E_Y E_{Y^*,S^*}\left(\sum_{t=1}^{T}\sum_{i=1}^{N} s_{ti}^* \log f_i(y_t^*; \hat{\beta}_i, \hat{\sigma}_i)\right)$$

$$= E_Y E_{Y^*,S^*}\left(\sum_{t=1}^{T}\sum_{i=1}^{N} s_{ti}^* \left(\log(2\pi\hat{\sigma}_i^2) + \frac{1}{\hat{\sigma}_i^2}(y_t^* - x_t'\hat{\beta}_i)^2\right)\right)$$

$$= E_Y \left(\sum_{i=1}^{N} T_i^* \log(2\pi\hat{\sigma}_i^2) + \sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} \sum_{t=1}^{T} E_{Y^*,S^*} \left(s_{ti}^* (y_t^* - x_t'\hat{\beta}_i)^2 \right) \right)$$

$$= E_Y \left(\sum_{i=1}^{N} T_i^* \log(2\pi\hat{\sigma}_i^2) + \sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} \sum_{t=1}^{T} E_{Y^*,S^*} \left(s_{ti}^* (\sigma_i^* \varepsilon_{ti} + x_t'\beta^* s_{ti}^* - x_t'\hat{\beta}_i)^2 \right) \right), \qquad \text{(A.2)}$$

where $T_i^* \equiv \pi_i^* T$, $\pi_i^* = E[s_{ti}^*]$ is the i-th element of the principal eigenvector of $P^* \pi^* = \pi^*$, and $\sigma_i^* \varepsilon_{ti} = y_t^* - x_t'\beta_i^*$. Moreover, from Hamilton (1990), we have $E_{Y^*,S^*}(s_{ti}^* \varepsilon_{ti} \mid x_t) = E_{Y^*}(\xi_{ti}^* \varepsilon_{ti} \mid x_t) = 0$.

Thus, the second term in (A.2) is

$$E_Y \left(\sum_{i=1}^{N} T_i^* \frac{(\sigma_i^*)^2}{\hat{\sigma}_i^2} + \sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} \sum_{t=1}^{T} x_t' E_{Y^*,S^*} \left(s_{ti}^* (\beta^* s_t^* - \hat{\beta}_i)(\beta^* s_t^* - \hat{\beta}_i)' \right) x_t \right)$$

$$= E_Y \left(\sum_{i=1}^{N} T_i^* \frac{(\sigma_i^*)^2}{\hat{\sigma}_i^2} + \sum_{i=1}^{N} \frac{1}{\hat{\sigma}_i^2} \pi_i^* \sum_{t=1}^{T} x_t' E_{Y^*} \left((\beta^* s_t^* - \hat{\beta}_i)(\beta^* s_t^* - \hat{\beta}_i)' \mid s_{ti}^* = 1 \right) x_t \right)$$

$$= E_Y \left(\sum_{i=1}^{N} T_i^* \frac{(\sigma_i^*)^2}{\hat{\sigma}_i^2} + \sum_{i=1}^{N} \pi_i^* \frac{(\hat{\beta}_i - \beta_i^*)' X'X (\hat{\beta}_i - \beta_i^*)}{\hat{\sigma}_i^2} \right). \qquad \text{(A.3)}$$

To evaluate (A.3), we first consider $(\hat{\beta}_i - \beta_i^*)' X'X (\hat{\beta}_i - \beta_i^*)$. Because $\hat{\beta}_i = (X'\hat{W}_i X)^{-1} X'\hat{W}_i y$ $= \beta_i^* + (X'\hat{W}_i X)^{-1} X'\hat{W}_i \varepsilon_i$, it follows that

$$(\hat{\beta}_i - \beta_i^*)' X'X (\hat{\beta}_i - \beta_i^*) = \varepsilon_i' \hat{W}_i X (X'\hat{W}_i X)^{-1} X'X (X'\hat{W}_i X)^{-1} X'\hat{W}_i \varepsilon_i$$

and

$$E_Y \left((\hat{\beta}_i - \beta_i^*)' X'X (\hat{\beta}_i - \beta_i^*) \right) = tr \left(E_Y [X(X'\hat{W}_i X)^{-1} X'X (X'\hat{W}_i X)^{-1} X'\hat{W}_i \varepsilon_i \varepsilon_i' \hat{W}_i] \right).$$

Moreover, $\hat{\xi}_{ti}$ is uncorrelated with x_t, and so $(X'\hat{W}_i X) = \sum_{t=1}^{T} \hat{\xi}_{ti} x_t x_t' \approx \hat{\pi}_i \sum_{t=1}^{T} x_t x_t' = \hat{\pi}_i (X'X)$, where $\hat{\pi}_i \equiv T^{-1} \sum_{t=1}^{T} \hat{\xi}_{ti}$. Thus,

$$E_Y \left((\hat{\beta}_i - \beta_i^*)' X'X (\hat{\beta}_i - \beta_i^*) \right) \approx tr \left(X(X'X)^{-1} X' E_Y (\hat{\pi}_i^{-2} \hat{W}_i \varepsilon_i \varepsilon_i' \hat{W}_i) \right).$$

Note that the diagonal and off-diagonals elements of $E_Y (\hat{\pi}_i^{-2} \hat{W}_i \varepsilon_i \varepsilon_i' \hat{W}_i)$ equal $E_Y (\varepsilon_{ti}^2 \hat{\xi}_{ti}^2 \hat{\pi}_i^{-2})$ and

$E_Y(\varepsilon_{ti}\varepsilon_{t-r,i}\hat{\xi}_{ti}\hat{\xi}_{t-r,i}\hat{\pi}_i^{-2})$, respectively.   Furthermore, $\hat{\pi}_i$ and $(\varepsilon_{ti},\hat{\xi}_{ti})$ are approximately

independent because $\hat{\pi}_i$ is the average of $\hat{\xi}_{ti}$ over t. Then, using this information and replacing

$\hat{\xi}_{ti}$ with s_{ti}^*, we obtain

$$E_Y\left(\varepsilon_{ti}^2\hat{\xi}_{ti}^2\hat{\pi}_i^{-2}\right) \approx E_Y\left(\varepsilon_{ti}^2\hat{\xi}_{ti}^2\right)E_Y\left(\hat{\pi}_i^{-2}\right)$$

$$\approx E_Y\left(\varepsilon_{ti}^2 s_{ti}^*\right)E_Y\left(\hat{\pi}_i^{-2}\right)$$

$$= \pi_i^*(\sigma_i^*)^2 E_Y\left(\hat{\pi}_i^{-2}\right),$$

and

$$E_Y\left(\varepsilon_{ti}\varepsilon_{t-r,i}\hat{\xi}_{ti}\hat{\xi}_{t-r,i}\hat{\pi}_i^{-2}\right) \approx E_Y\left(\varepsilon_{ti}\varepsilon_{t-r,i}s_{ti}^*s_{t-r,i}^*\right)E_Y\left(\hat{\pi}_i^{-2}\right)$$

$$= E_Y\left(E_Y\left(\varepsilon_{ti}\mid\varepsilon_{t-r,i},s_{ti}^*,s_{t-r,i}^*\right)\varepsilon_{t-r,i}s_{ti}^*s_{t-r,i}^*\right)E_Y\left(\hat{\pi}_i^{-2}\right)$$

$$= 0.$$

Consequently, we have

$$E_Y\left((\hat{\beta}_i - \beta_i^*)'X'X(\hat{\beta}_i - \beta_i^*)\right) \approx \pi_i^*(\sigma_i^*)^2 E_Y\left(\hat{\pi}_i^{-2}\right)tr\left(X(X'X)^{-1}X'\right)$$

$$\equiv (\pi_i^*)^{-1}(\sigma_i^*)^2 \lambda_i K , \tag{A.4}$$

where $\lambda_i \equiv (\pi_i^*)^2 E_Y(\hat{\pi}_i^{-2})$.

Next consider $\hat{\sigma}_i^2$ in (A.3). Since $y - X\hat{\beta}_i = y - X(X'\hat{W}_iX)^{-1}X'\hat{W}_i(X\beta_i^* + \varepsilon_i)$

$= (I - X(X'\hat{W}_iX)^{-1}X'\hat{W}_i)\varepsilon_i$, we have

$$\hat{\sigma}_i^2 = \hat{T}_i^{-1}(y - X\hat{\beta}_i)'\hat{W}_i(y - X\hat{\beta}_i)$$

$$= \hat{T}_i^{-1}\varepsilon_i'\left(I - \hat{W}_iX\left(X'\hat{W}_iX\right)^{-1}X'\right)\hat{W}_i\varepsilon_i$$

$$= (\hat{\pi}_iT)^{-1}\varepsilon_i'\hat{W}_i\varepsilon_i - (\hat{\pi}_iT)^{-1}\varepsilon_i'\hat{W}_iX\left(X'\hat{W}_iX\right)^{-1}X'\hat{W}_i\varepsilon_i.$$

Moreover, approximating $(X'\hat{W}_iX)$ with $\hat{\pi}_i(X'X)$ as above, we obtain

26

$$E_Y(\hat{\sigma}_i^2) \approx T^{-1}E_Y\left(\hat{\pi}_i^{-1}\varepsilon_i'\hat{W}_i\varepsilon_i\right) - T^{-1}E_Y\left(\hat{\pi}_i^{-2}\varepsilon_i'\hat{W}_iX(X'X)^{-1}X'\hat{W}_i\varepsilon_i\right)$$

$$= E_Y\left(\varepsilon_{ti}^2\hat{\xi}_{ti}\hat{\pi}_i^{-1}\right) - T^{-1}tr\left(X(X'X)^{-1}X'E_Y\left(\hat{W}_i\varepsilon_i\varepsilon_i'\hat{W}_i\hat{\pi}_i^{-2}\right)\right)$$

$$\approx E_Y\left(\varepsilon_{ti}^2\hat{\xi}_{ti}\hat{\pi}_i^{-1}\right) - T^{-1}(\pi_i^*)^{-1}(\sigma_i^*)^2\lambda_iK .$$

Because $\hat{\pi}_i$ and $(\varepsilon_{ti}, \hat{\xi}_{ti})$ are approximately independent and $E_Y(\varepsilon_{ti}^2\hat{\xi}_{ti}) \approx E_Y(\varepsilon_{ti}^2\xi_{ti}^*) = \pi_i^*(\sigma_i^*)^2$,

we have $E_Y(\varepsilon_{ti}^2\hat{\xi}_{ti}\hat{\pi}_i^{-1}) \approx E_Y(\varepsilon_{ti}^2\xi_{ti}^*)E_Y(\hat{\pi}_i^{-1}) = \pi_i^*(\sigma_i^*)^2 \, E_Y(\hat{\pi}_i^{-1})$. Consequently,

$$E_Y\left(\frac{T_i^*\hat{\sigma}_i^2}{(\sigma_i^*)^2}\right) \approx \delta_iT_i^* - \lambda_iK , \tag{A.5}$$

where $\delta_i \equiv \pi_i^*E_Y(\hat{\pi}_i^{-1})$.

Using (A.4) and (A.5) together with the Chi-squared approximation to $\dfrac{T_i^*\hat{\sigma}_i^2}{(\sigma_i^*)^2}$ proposed

by Cleveland and Devlin (1988), we obtain

$$E_Y\left(\frac{(\sigma_i^*)^2}{T_i^*\hat{\sigma}_i^2}\right) \approx \frac{1}{\delta_iT_i^* - \lambda_iK - 2} \tag{A.6}$$

and

$$E_Y\left(\frac{(\hat{\beta}_i - \beta_i^*)'X'X(\hat{\beta}_i - \beta_i^*)}{\hat{\sigma}_i^2}\right) = T_i^*E_Y\left(\frac{(\hat{\beta}_i - \beta_i^*)'X'X(\hat{\beta}_i - \beta_i^*)/(\sigma_i^*)^2}{T_i^*\hat{\sigma}_i^2/(\sigma_i^*)^2}\right)$$

$$\approx \frac{T_i^*(\pi_i^*)^{-1}\lambda_iK}{\delta_iT_i^* - \lambda_iK - 2} . \tag{A.7}$$

Substituting (A.6) and (A.7) into (A.3) in conjunction with (A.2), we find that the average KL

information loss in (A.1) is

$$\tilde{d}_{KL} \approx \sum_{i=1}^{N}T_i^*E_Y\left(\log 2\pi\hat{\sigma}_i^2\right) + \sum_{i=1}^{N}\frac{(T_i^*)^2}{\delta_iT_i^* - \lambda_iK - 2} + \sum_{i=1}^{N}\frac{T_i^*\lambda_iK}{\delta_iT_i^* - \lambda_iK - 2} - 2E_Y\left(E_{Y^*}\left(R(N,Y^*,\hat{\theta})\right)\right).$$

Finally, replacing T_i^*, $E_Y(\log 2\pi\hat{\sigma}_i^2)$ and $E_Y\left(E_{Y^*}(R(N,Y^*,\hat{\theta}))\right)$ with their in-sample estimates,

we obtain an estimate of $\tilde{d}_{KL}$:

$$MSC = \sum_{i=1}^{N} \hat{T}_i \log 2\pi\hat{\sigma}_i^2 + \sum_{i=1}^{N} \frac{\hat{T}_i(\hat{T}_i + \lambda_i K)}{\delta_i \hat{T}_i - \lambda_i K - 2} - 2R(N, Y, \hat{\theta})$$

$$= -2\log(f(Y, \hat{\theta})) + \sum_{i=1}^{N} \frac{\hat{T}_i(\hat{T}_i + \lambda_i K)}{\delta_i \hat{T}_i - \lambda_i K - 2},$$

where $\log f(Y; \hat{\theta}) = -\frac{1}{2} \sum_{t=1}^{T} \hat{T}_i (\log 2\pi\hat{\sigma}_i^2 + 1) + R(N, Y, \hat{\theta})$.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In 2nd International Symposium on Information Theory, B.N. Petrov and F. Csaki (Eds.), 267-281. Budapest: Akademia Kiado.

Akaike, H., 1985. Prediction and entropy. In A Celebration of Statistics. A.C. Atkinson and S.E. Fienberg (Eds.), 1-24. New York: Springer.

Ang, A., Bekaert, G., 2002a, Regime switches in interest rates. Journal of Business and Economic Statistics 20, 163-182.

Ang, A., Bekaert, G., 2002b, International asset allocation with regime shifts. Review of Financial Studies 15, 1137-1187.

Bates, J. M., Granger, C. W. J. 1969. The combination of forecasts. Operations Research Quarterly 20, 451-468.

Baum, L.E., Petrie, T., 1966. Statistical inference for probabilistic functions of finite state Markov chains. The Annals Mathematical Statistics 37, 1554-1563.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics 41, 164-171.

Bickel, P.J., Ritov, Y., Rydén, T., 1998. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. Annals of Statistics 26, 1614-1635.

Bucklin, R.E., Gupta, S., 1999. Commercial use of UPC scanner data: industry and academic perspectives. Marketing Science 18, 247-273.

Bunke, H., Caelli, T., (Eds.), 2001. Hidden Markov models: applications in computer vision. World Scientific, Series in Machine Perception and Artificial Intelligence 45.

Burnham, K. P. and Anderson, D. R. 2004. Multimodel Inference: Understanding AIC and BIC in model selection. Sociological Methods and Research 33, 2, 261-304.

Burnham, K.P. Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach (2nd Ed). New York: Springer.

Cleveland, W.S., Devlin, S.J., 1988. Locally weighted regression: an approach to regression analysis by local fitting. Journal of the American Statistical Association 83, 596-610.

deLeeuw, Jan. 1992. Introduction to Akaike (1973) information theory and extension of the maximum likelihood principle. In Breakthroughs in Statistics, Vol. 1, Edited by Samuel Kotz and Norman Johnson, p. 599-609. London: Springer-Verlag.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society (Series B) 39, 1-38.

Durbin, R., Eddy, S., Krogh, A., Mitchison, G., 1998. Biological Sequence Analysis. Cambridge University Press.

Engel, C., Hamilton, J.D., 1990, Long swings in the dollar: are they in the data and do markets know it? American Economic Review 80, 689-713.

Feichtinger, G., Hartl, R.F., Sethi, S.P., 1994. Dynamic optimal control models in advertising: recent developments. Management Science 40, 195-226.

Gabriel, K.R., 1959. The distribution of the number of successes in a sequence of dependent trials. Biometrika 46, 454-460.

Garcia, R., 1998. Asymptotic null distribution of the likelihood ratio test in Markov switching models. International Economic Review 39, 763-788.

Garcia, R., Perron, P., 1996. An analysis of the real interest rate under regime shifts. Review of

Economics and Statistics 78, 111-125.

Ghysels, E., McCulloch, R.E., and Tsay, R.S., 1998. Bayesian inference for periodic regime-switching models. Journal of Applied Econometrics 13, 129-143.

Hamilton, J.D, Susmel, R., 1994. Autoregressive conditional heteroskedasticity and changes in regime. Journal of Econometrics 64, 307-333.

Hamilton, J.D., 1989. A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica 57, 357-384.

Hamilton, J.D., 1990. Analysis of time series subject to changes in regime. Journal of Econometrics 45, 39-70.

Hansen, B.E., 1992. The likelihood ratio test under nonstandard conditions: testing the Markov switching model of GNP. Journal of Applied Econometrics 7, S61-82.

Hartigan, J.A., 1977. Distribution problems in clustering. In Classification and Clustering. J. van Ryzin (Ed). New York, Academic Press.

Hathaway, R. J., 1985. A constraint formulation of maximum-likelihood estimation for normal mixture distributions. Annals of Statistics 13, 795-800.

Hjort, N. L., Claeskens, G. 2003. Frequentist model average estimators. Journal of the American Statistical Society 98, 464, 879-899.

Hoeting, J. A., Madigan, D., Raftery, A., Volinsky, C. T. 1999. Bayesian model averaging: a tutorial. Statistical Science 14, 4, 382-417.

Horowitz, J.L., 1998. Semiparametric Methods in Econometrics. New York: Springer.

Hurvich, C.M., Tsai, C.L., 1989. Regression and time series model selection in small samples. Biometrika 76, 297-307.

Kass, R. E., Raftery, A., 1995. Bayes factors. Journal of the American Statistical Association 90, 773-795.

Kim, C.-J., 1994. Dynamic linear models with Markov switching. Journal of Econometrics 60, 1-22.

Kim, C.-J., Nelson, C.R., 1999a. Has the U.S. economy become more stable? A bayesian approach based on a Markov-switching model of the business cycle. Review of Economics and Statistics 81, 608-616.

Kim, C.-J., Nelson, C. R., 1999b. State-Space Models with Regime Switching. Cambridge, MA: MIT Press.

Kullback, S., Leibler, R. A., 1951. On information and sufficiency. The Annals of Mathematical Statistics 22, 1, 79-86.

Leamer, E. E., 1978. Specification Searches: Ad Hoc Inference with Non-experimental Data. New York, NY: Wiley.

Leroux, B., 1992. Consistent estimation of a mixing distribution. Annals of Statistics 20, 1350-1360.

Leroux, B.G., Puterman, M.L., 1992. Maximum-penalized-likelihood estimation for independent and Markov-dependent mixture models. Biometrics 48, 545-558.

Linhart, H., Zucchini, W., 1986. Model Selection. New York, NY: Wiley.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 281-297. Berkeley, Calif.: University of California Press.

Mantrala, M.K., 2002. Allocating marketing resources. In Handbook of Marketing, B.A. Weitz and R. Wensley (Eds.). Sage Publications.

McConnell, M. M., Perez-Quiros, G., 2000. Output fluctuations in the United States: what has

changed since the early 1980's? American Economic Review 90, 1464-1476.

McLachlan, G. J, Peel, D., 2000. Finite Mixture Models. New York: Wiley.

McLachlan, G. J. and Krishnan, T. 1997. The EM Algorithm and Extensions. New York, NY: John Wiley & Sons.

Naik, P. A., Mantrala, M. K., Sawyer, A. G., 1998. Planning media schedules in the presence of dynamic advertising quality. Marketing Science 17, 3, 214-235.

Naik, P.A., Raman, K., 2003. Understanding the impact of synergy in multimedia communications. Journal of Marketing Research 40, 375-388.

Naik, P.A., Tsai, C.L., 2001. Single-index model selections. Biometrika 88, 821-832.

Palda, K.S., 1964. The Measurement of Cumulative Advertising Effects, Englewood Cliffs, New Jersey: Prentice Hall.

Psaradakis, Z., Sola, M., 1998. Finite-sample properties of the maximum likelihood estimator in autoregressive models with Markov switching. Journal of Econometrics 86, 369-386.

Psaradkis, Z., Spagnolo, N., 2003. On the determination of the number of regimes in Markov-switching autoregressive models. Journal of Time Series Analysis 24, 2, 237-252.

Rabiner, L., Juang, B.H., 1993. Fundamentals of Speech Recognition. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Raftery, A. E., 1996. Approximate Bayes factors and accounting for model uncertainty in generalized linear models. Biometrika 83, 2, 251-266.

Rényi, A. 1970. Foundations of Probability. San Francisco, CA: Holden-Dat, Inc.

Sawa, T., 1978. Information criteria for discriminating among alternative regression models. Econometrica 46, 1273-1291.

Shannon, C. E., 1948. A mathematical theory of communication. Bell System Technical Journal 27, 379-423.

Sin, C.Y., White, H., 1996. Information criteria for selecting possibly misspecified parametric models. Journal of Econometrics 71, 207-225.

Stock, J.H., Watson, M.W., 2003. Has the business cycle changed? Evidence and explanations. Working Paper, http://www.wws.princeton.edu/~mwatson/papers/jh_2.pdf.

Timmermann, A., 2001. Structural breaks, incomplete information and stock prices. Journal of Business and Economic Statistics 19, 299-315.

Timmermann, A., 2005. Forecast combinations. In Handbook of Economic Forecasting (Edited by Elliott, G., Granger, C. W. J., and Timmermann, A.) North Holland.

Yang, Y., 2001. Adaptive regression by mixing. Journal of the American Statistical Association 96, 454, 574-588.

Zellner, A. 2002. Information processing and Bayesian analysis. Journal of Econometrics 107, 41-50.

Figure 1: Proximity of MSC to the KL distance

Panel A. $K = 3$, Various N

Panel B. $N = 2$, Various K
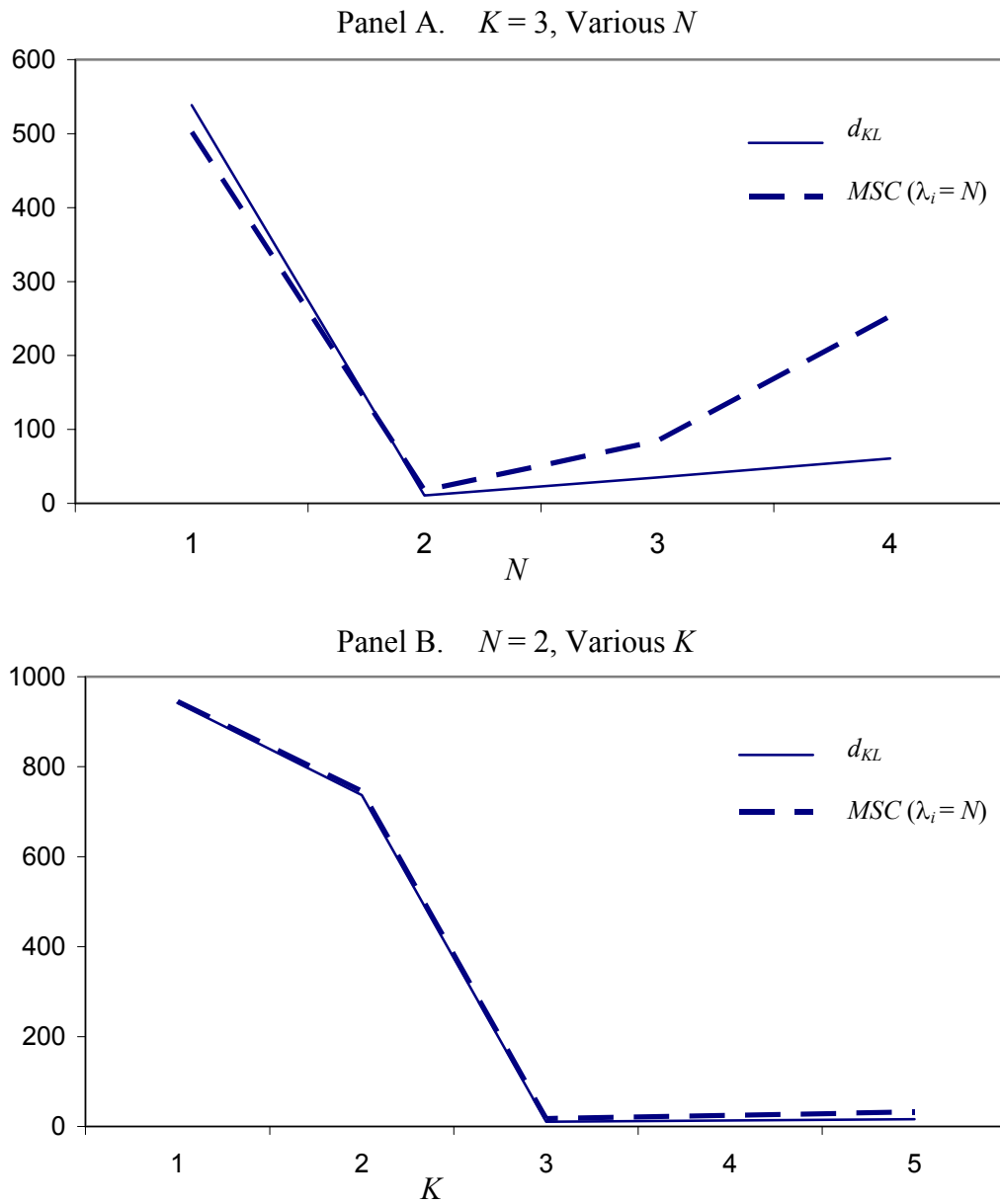
Figure 2: Smoothed state probabilities for GNP growth model

Panel A: Recession

Panel B: Expansion (type one)

Panel C: Expansion (type two)
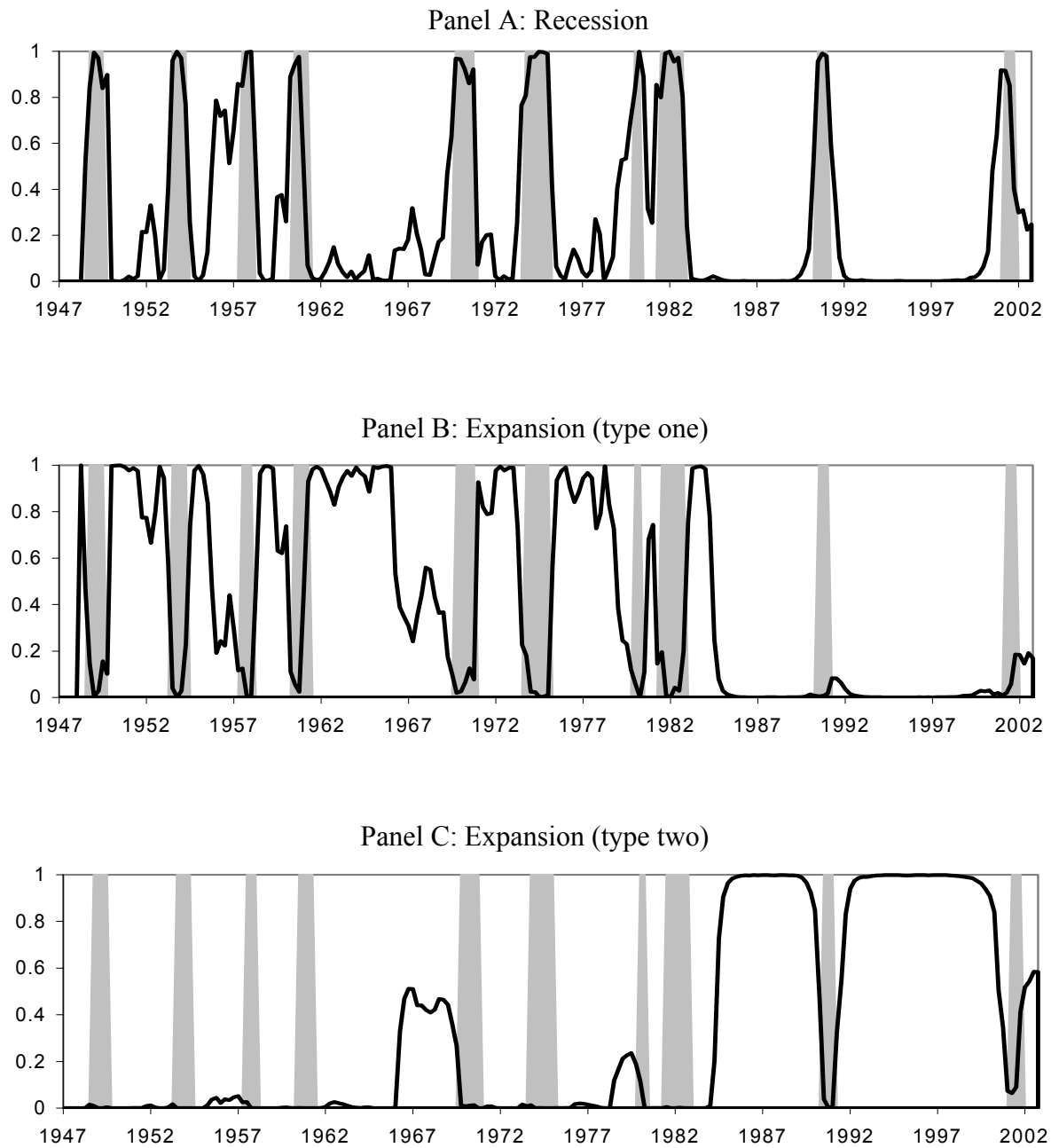
NBER recession dates are indicated by gray bars.

Table 1. Joint selection frequency in 1000 realizations (fixed regressors)

Panel A. $MSC_{\lambda = 1}$

Variables (K)	States (N)				column sum
	1	2	3	4	
1	0	0	0	0	0
2	0	0	0	0	0
$K^0 = 3$	0	309	190	167	666
4	0	29	85	97	211
5	0	22	47	54	123
row sum	0	360	322	318	1000

Panel B. $MSC_{\lambda = N}$

Variables (K)	States (N)				column sum
	1	$N^0 = 2$	3	4	
1	0	0	0	0	0
2	0	0	0	0	0
$K^0 = 3$	0	992	0	0	992
4	0	8	0	0	8
5	0	0	0	0	0
row sum	0	1000	0	0	1000

Panel C. $MSC_{\lambda = N^2}$

Variables (K)	States (N)				column sum
	1	$N^0 = 2$	3	4	
1	0	0	0	0	0
2	0	0	0	0	0
$K^0 = 3$	0	1000	0	0	1000
4	0	0	0	0	0
5	0	0	0	0	0
row sum	0	1000	0	0	1000

Table 2. Frequency of correctly selecting both states and variables in 1000 realizations (fixed regressors)

	$\text{MSC}_{\lambda=1}$	$\text{MSC}_{\lambda=N}$	$\text{MSC}_{\lambda=N^2}$
Large sample ($T = 250$)			
Low noise ($\sigma^0 = 0.5$)	309	992	1000
High noise ($\sigma^0 = 1$)	124	981	998
Small sample ($T = 100$)			
Low noise ($\sigma^0 = 0.5$)	521	951	861
High noise ($\sigma^0 = 1$)	393	907	741

Table 3. Joint selection frequency in 1000 realizations (autoregression)

Panel A. $MSC_{\lambda=1}$

Variables (K)	1	States (N) $N^0 = 2$	3	4	column sum
1	0	0	0	0	0
2	0	0	0	1	1
$K^0 = 3$	0	524	169	78	771
4	0	52	62	36	150
5	0	26	33	19	78
row sum	0	602	264	133	1000

Panel B. $MSC_{\lambda=N}$

Variables (K)	1	States (N) $N^0 = 2$	3	4	column sum
1	0	0	0	0	0
2	0	3	0	0	3
$K^0 = 3$	0	979	1	0	980
4	0	16	0	0	16
5	0	1	0	0	1
row sum	0	999	1	0	1000

Panel C. $MSC_{\lambda=N^2}$

Variables (K)	1	States (N) $N^0 = 2$	3	4	column sum
1	0	0	0	0	0
2	0	14	0	0	14
$K^0 = 3$	0	984	0	0	984
4	0	2	0	0	2
5	0	0	0	0	0
row sum	0	1000	0	0	1000

Table 4. Frequency of correctly selecting both states and variables in 1000 realizations (autoregression)

	$MSC_{\lambda=1}$	$MSC_{\lambda=N}$	$MSC_{\lambda=N^2}$
Large sample ($T = 250$)			
Low noise ($\sigma^0 = 0.5$)	524	979	984
High noise ($\sigma^0 = 1$)	264	974	785
Small sample ($T = 100$)			
Low noise ($\sigma^0 = 0.5$)	648	744	171
High noise ($\sigma^0 = 1$)	389	402	1

Table 5. Frequency of correctly selecting a single-state model in 1000 realizations

	$MSC_{\lambda=1}$	$MSC_{\lambda=N}$	$MSC_{\lambda=N^2}$
Panel A. Fixed Regressors			
Large sample ($T = 250$)			
Low noise ($\sigma^0 = 0.5$)	92	992	999
High noise ($\sigma^0 = 1$)	38	991	1000
Small sample ($T = 100$)			
Low noise ($\sigma^0 = 0.5$)	204	994	997
High noise ($\sigma^0 = 1$)	258	998	1000
Panel B. Autoregression			
Large sample ($T = 250$)			
Low noise ($\sigma^0 = 0.5$)	96	980	1000
High noise ($\sigma^0 = 1$)	51	973	1000
Small sample ($T = 100$)			
Low noise ($\sigma^0 = 0.5$)	207	945	998
High noise ($\sigma^0 = 1$)	78	923	999

Table 6. Joint selection frequency in 1000 realizations by AIC

Variables (K)	1	States (N) $N^0 = 2$	3	4	column sum
1	0	0	0	0	0
2	0	0	0	0	0
$K^0 = 3$	0	481	106	30	617
4	0	66	76	48	190
5	0	34	97	62	193
row sum	0	581	279	140	1000

Table 7: Estimated parameters for the three-state model for the U.S. GNP growth

Parameters for each state i	State 1	State 2	State 3
Mean Growth Rate, $\hat{\beta}_i$	–0.10 (0.29)	1.50 (0.18)	0.85 (0.06)
Noise level, $\hat{\sigma}_i$	0.95 (0.15)	0.91 (0.09)	0.42 (0.04)
Transition Probability Matrix, $\hat{P}$			
$\Pr(s_{ti}=1\mid s_{t-1,1}=1)$	0.78 (0.11)	0.20 (0.10)	0.02 (0.04)
$\Pr(s_{ti}=1\mid s_{t-1,2}=1)$	0.12 (0.08)	0.85 (0.08)	0.03 (0.02)
$\Pr(s_{ti}=1\mid s_{t-1,3}=1)$	0.04 (0.05)	0.00 (0.07)	0.96 (0.05)

Standard errors (in parentheses) were computed from the outer product of scores.

Table 8: Estimated parameters for the two-state model for media advertising

Parameters for each state i	State 1	State 2
Advertising effectiveness $\hat{\beta}_i^{(1)}$	0.43 (0.17)	1.17 (0.23)
Carryover effect, $\hat{\beta}_i^{(2)}$	0.53 (0.14)	0.27 (0.09)
Intercept, $\hat{\beta}_i^{(0)}$	1.05 (0.45)	0.26 (0.20)
Noise level, $\hat{\sigma}_i$	0.53 (0.09)	0.10 (0.02)
Transition Probability Matrix, $\hat{P}$		
$\Pr(s_{ti}=1 \mid s_{t-1,1}=1)$	0.96 (0.04)	0.04 (0.04)
$\Pr(s_{ti}=1 \mid s_{t-1,2}=1)$	0.06 (0.07)	0.94 (0.07)

Standard errors (in parentheses) were computed from the outer product of scores.